Nucleotide sequence of an unusual regionally expressed silkmoth chorion RNA: Predicted primary and secondary structures of an architectural protein

(chorion morphogenesis/chorion gene evolution/consensus sequence)

JEROME C. REGIER AND PAUL PACHOLSKI

Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, IL 60201

Communicated by Fotis C. Kafatos, May 2, 1985

ABSTRACT We have sequenced DNA clones corresponding to the entire coding and 3' untranslated regions and almost all of the 3' untranslated region of a silkmoth chorion RNA which is expressed largely in a subpopulation of follicular epithelial cells (aeropyle crown region). This RNA encodes the E1 protein, one of two components of the prominent "filler" that helps mold the shape of aeropyle crowns. The conceptually translated E1 sequence reveals an alternation in hydrophobic and hydrophilic stretches of amino acids that correlates with certain predictions about its secondary structure. E1 is unusual in revealing no sequence homology with other known chorion sequences and in having an unusually long 3' untranslated region. Sequence analysis of the 5' end of the E1 gene has identified an intron near the end of the signal peptide-encoding region, a feature shared with other chorion genes.

Choriogenesis in silkmoth includes the sequential synthesis of more than 100 distinct proteins during its 2-day duration (1). During the first nine-tenths of choriogenesis, all cells in the chorion-producing follicular epithelium produce the same proteins at the same times and at the same rates (2). These proteins are secreted and extracellularly to form a largely lamellar structure (3). Then, during the "very late" period, two major subpopulations of follicle cells become distinguishable on the basis of distinct patterns of chorion protein and RNA synthesis (ref. 2 and unpublished observations). Cells in the aeropyle crown region begin synthesizing a distinct subset of chorion proteins that assemble into prominent crown-shaped surface structures underlying each three-cell junction. At the ultrastructural level, aeropyle crowns consist of several lamellae similar to, but thinner than, those of the underlying lamellar chorion (4). Within each aeropyle crown is a ball of spongy material called "filler," which appears to have an architectural role and extends into the underlying narrow aeropyle channel (2). Filler consists of only two distinct proteins, called E1 and E2, whereas lamellae, including those of the aeropyle crowns, are biochemically much more complex and include most members of the A-, B-, and C-size classes of chorion proteins (2). The other major population of follicle cells is found in the so-called flat region. Very late-period-specific proteins appear to be synthesized in this region as well, although their absolute rates are reduced by a factor of 10 or more relative to cells in the aeropyle crown region. In the flat region, no aeropyle crowns ever form and filler is confined to the aeropyle channels.

Regier et al. (5) have described the molecular cloning of very late-period-specific nucleic acid sequences that encode the filler proteins, E1 and E2. Our interest in further characterization of E sequences is threefold: (i) to search for features shared between very late period E sequences and previously characterized chorion sequences that might indicate evolutionary relatedness, (ii) to begin to understand how E1 and E2 sequences assemble to form filler, and (iii) to describe the molecular basis of regionalization, including cell-specific expression at the RNA level and organization of regionally expressed genes. Concerning the first goal, it should be noted that the extensive chorion sequence data base (more than 20 partial or complete sequences) includes lamellar components only (6-8). Within this group, striking homologies have been found for each of the five chorion multigene families characterized. Homologies are apparent even between different gene families, so that all lamellar chorion sequences to date may be placed into a chorion "superfamily" (6, 9). In this report we present the complete E1 amino acid sequence as deduced from its encoding RNA, and we compare its sequence with those of lamellar chorion proteins.

MATERIALS AND METHODS

Isolation and Sequencing of Clones. The construction of an Antheraea polyphemus silkmoth chorion cDNA library in the plasmid pBR322 and the isolation and characterization of E1-encoding cDNA clones have been described (5). To facilitate sequencing by the chemical degradation method (10), we isolated intact plasmid inserts from two of these clones and reinserted them into the plasmid pUC8 (11). Isolation of the E1 gene-containing chromosomal clone AP8 will be described in a future publication.

RESULTS

Sequencing Strategy. In a previous report, the cDNA clone pcvl 3 was shown to encode the E1 protein and to be near full-length (5). The strategy used to sequence the complete 752-base-pair (bp) insert [minus oligo(G-C) tails] is shown in Fig. 1. Of the six possible translational reading frames (from each strand), the correct one was easily identifiable based on the absence of translational termination codons over a significant stretch and by comparison of the published E1 amino acid composition with that deduced from the clone (12). After conceptual translation of the pcvl 3 sequence, it was concluded that the clone had an incomplete 5' end. To extend the 5' end of the sequence, we resequenced the original cDNA library for additional E1 clones. pcvl 17 was chosen and sequenced from both ends. It contains 45 additional base pairs at its 5' end and 44 fewer at its 3' end. Conceptual translation of pcvl 17 suggested that its 5' end terminated within the signal peptide sequence. Therefore, we isolated and sequenced the 5' end of the E1 gene found within chromosomal clone AP8. Alignment of the AP8 and pcvl 17

Abbreviation: bp, base pair(s).
overlapping clones that contain sequences encoding the E1 chorion protein. The inserts of the cDNA clones pcvl 3 and pcvl 17 were sequenced from restriction sites present within the pUC8 polynucleotides. pcvl 3 was also present from the internal sites shown. The 5' end of the E1 gene present within chromosomal clone AP8 was sequenced in both directions from a HincII site present within an intron (indicated by a triangle). Arrows identify the direction and extent of DNA sequence determination. Length is given in bp, and +1 is the position of the first translated nucleotide.

sequences revealed near identity except for an ~200-bp insertion in the gene between nucleotides 45 and 46 (Fig. 2). We consider this insertion to be an intron for two reasons: (i) sequence identity between AP8 and pcvl 17 extends on both sides of the intron and (ii) the borders of the E1 intronic sequences (GTAAGT....TGTGCTGAG) closely resemble those of other introns (15). Using the E1 gene sequence, we extended the 5' end 16 bp until we came to an ATG codon, representing the translational initiation site. An additional in-frame ATG sequence is found 24 bp upstream, but it is followed by an in-frame termination codon. No additional in-frame ATG sequences are found within the next 167 bp of the genome.

The presence of a putative "TATA box" at nucleotides −49 to −43 provides evidence that the proposed site of translational initiation is close to the 5' end of the RNA transcript. In turn, the location of the TATA box suggests that the site of transcriptional initiation is located somewhere between nucleotides −25 and −21 (box above sequence in Fig. 2). This site encompasses one of a pair of similar octanucleotides (nucleotides −23 to −16 and −14 to −7).

E1 Sequence. By aligning the three clone sequences as shown in Fig. 1, we can construct an 863-bp hybrid sequence that spans the entire E1 protein coding region as well as most of the mature E1 RNA coding region (Fig. 2). Is this hybrid sequence derived from only one gene? Southern hybridization analysis of genomic DNA and of E1 genomic clones indicates a small number of E1 genes per haploid genome, perhaps one but certainly no more than four (unpublished data). Whether the distinct E1 sequences are alleles or

FIG. 1. Restriction maps and sequencing strategies for three overlapping clones that contain sequences encoding the E1 chorion protein. The inserts of the cDNA clones pcvl 3 and pcvl 17 were sequenced from restriction sites present within the pUC8 polynucleotides. pcvl 3 was also present from the internal sites shown. The 5' end of the E1 gene present within chromosomal clone AP8 was sequenced in both directions from a HincII site present within an intron (indicated by a triangle). Arrows identify the direction and extent of DNA sequence determination. Length is given in bp, and +1 is the position of the first translated nucleotide.

Fig. 2. E1 nucleotide sequence and its conceptually translated polypeptide. Three overlapping E1 clones were sequenced partially or in their entirety (see Fig. 1). Identical overlapping sequences are indicated by dots. Substitutions are shown as nucleotides interrupting the dotted line. pcvl 17 contains a single deletion and a single insertion relative to pcvl 3 (identified by parentheses at nucleotides 605–609 and between 759 and 760). Nucleotide differences result in amino acid differences at two positions (residues 34 and 46). A thick line identifies a "TATA box" at nucleotides −49 to −43. Another thick line at nucleotides −15 to −20 identifies a hexanucleotide sequence found in 5' untranslated regions of other chorion genes. Thin lines identify short repeated sequences. A box above nucleotides −24 to −16 identifies a region that contains the site of transcriptional initiation, based on distance from TATA box in a large number of cases. Dots within the box identify the four nucleotides most likely to represent the initiation site, based on similarities in sequence with other chorion genes (8, 13, 14). The solid triangle identifies the location of an intron (I). At nucleotide position 669, n signifies undetermined. The three most probable sites separating signal peptide from mature protein are each identified by an x. Asterisks beneath amino acid residues within the tetradecameric repeats represent the two conservative substitutions.
not, hybridization analysis shows that they are very similar in primary structure to each other.

The conceptually translated sequence of 169 amino acid residues preceding the termination codon is shown below its encoding nucleotide sequence (Fig. 2). Surprisingly, 302 nucleotides follow the termination codon, significantly more than in previously described chorion sequences (8, 13). Within this 3' untranslated region are interspersed multiple termination codons (8–11, depending on the reading frame) and four possible AATAAA polyadenylation signals (16). The 3' untranslated region must be nearly complete, since we have previously estimated the E1 RNA to be 850 nucleotides long, and the approximately 830-nucleotide transcribed sequence in Fig. 2 does not include a poly(A) tail (5).

E1, like other chorion proteins, is secreted extracellularly and presumably is synthesized with an amino-terminal signal peptide (17, 18). We have identified the three positions (each marked by an x in Fig. 2) most likely to separate signal peptide sequence from the mature protein sequence based on the observation that the last residue in a signal peptide is small and neutral and that the resulting conceptualized amino acid composition should closely match the published values (12, 19). The polypeptide defined by residues 17–169 is slightly favored on the basis of composition (Table 1). Slight differences between the two amino acid compositions are easily ascribed to population polymorphism and to minor inaccuracies inherent in amino acid analysis. This 153-residue polypeptide has a molecular weight of 15,694, making it equal in length to but slightly larger in molecular weight than the largest known class B protein (13). This is consistent with their respective mobilities on NaDodSO4/polyacrylamide gels, where E1 protein migrates slightly more slowly than the B protein (5).

A primary structural feature characteristic of other chorion protein sequences is the presence of short internal repeats (20). E1 contains a single direct tetradecapeptide repeat (residues 114–127 and 128–141) with only two conservative differences. Other than this no stretches longer than three amino acid residues are present more than once in the sequence.

Table 1. Comparison of amino acid composition of purified E1 protein and that deduced from cloned cDNA

<table>
<thead>
<tr>
<th>Residues per 100 residues</th>
<th>E1 protein*</th>
<th>cDNA clone†</th>
<th>Column 1 – column 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cys</td>
<td>1.7</td>
<td>2.6</td>
<td>-0.9</td>
</tr>
<tr>
<td>Asx</td>
<td>6.8</td>
<td>6.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Thr</td>
<td>10.0</td>
<td>10.5</td>
<td>-0.5</td>
</tr>
<tr>
<td>Ser</td>
<td>8.4</td>
<td>9.8</td>
<td>-1.4</td>
</tr>
<tr>
<td>G1x</td>
<td>12.0</td>
<td>11.1</td>
<td>0.9</td>
</tr>
<tr>
<td>Pro</td>
<td>4.9</td>
<td>4.6</td>
<td>0.3</td>
</tr>
<tr>
<td>Gly</td>
<td>12.2</td>
<td>11.1</td>
<td>1.1</td>
</tr>
<tr>
<td>Ala</td>
<td>10.0</td>
<td>9.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Val</td>
<td>8.6</td>
<td>8.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Met</td>
<td>1.2</td>
<td>1.3</td>
<td>-0.1</td>
</tr>
<tr>
<td>Ile</td>
<td>4.0</td>
<td>3.9</td>
<td>0.1</td>
</tr>
<tr>
<td>Leu</td>
<td>8.5</td>
<td>8.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Tyr</td>
<td>2.3</td>
<td>2.6</td>
<td>-0.3</td>
</tr>
<tr>
<td>Phe</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>His</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Lys</td>
<td>5.8</td>
<td>5.9</td>
<td>-0.1</td>
</tr>
<tr>
<td>Trp</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Arg</td>
<td>3.4</td>
<td>3.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*From ref. 12.
†Includes residues 17–169 of the conceptually translated hybrid sequence shown in Fig. 2, with Asn at residue 34 and Cys at residue 46.

Comparison with Other Sequences. We have searched for similarities between E1 and other published chorion sequences, using a homology matrix computer program (21). Sequences chosen for comparison were representatives from each of the five chorion multigene families characterized to date (A, B, C, Hc-A, and Hc-B; see ref. 6 and 9). No obvious similarities have been found in either the coding or 3' untranslated regions. An example, Fig. 3 shows that a typical B-family protein shares no similarity in sequence with E1, whereas the same B-family protein shares obvious similarities with a typical A-family protein. To our knowledge, E1 represents the first chorion sequence that displays no similarities with other chorion sequences in their coding or 3' untranslated regions. We have also compared E1 with all sequences in the National Institutes of Health DNA sequence library, the European Molecular Biology Organization Nucleotide Sequence Data Library, and the National Biomedical Research Foundation Protein Sequence Database. Again, no obvious similarities were observed.

A comparison of the short (20–35 nucleotides) transcribed but untranslated region at the 5' ends of chorion genes reveals a similar hexanucleotide sequence in all cases (Fig. 4 and Discussion).

Hydropobicity and Secondary Structure Prediction. Kyte and Doolittle (22) described a simple computer program which progressively evaluates the hydrophilicity and hydrophobicity of a protein along its amino acid sequence as a set of hydrophathy indices. What is strikingly apparent across the E1 sequence is that the hydrophathy index is not uniform (Fig. 5, top panel). Instead, there are five prominent peaks of

![Fig. 3. Homology matrix plots of a typical B-family chorion protein sequence (called 401, see ref. 20) with E1 (upper plot) and with a typical A-family chorion protein sequence (called 18, lower plot; see ref. 20). Sequentially overlapping undecapeptides from pairs of sequences were compared for degrees of similarity. All comparisons having at least 62% identity are plotted (A = 99–100% match, B = 97–98% match, etc.). A scale factor of 0.90 was used to weight residues within undecapeptides (21).](image-url)
hydrophobicity, the first four (from left to right) of which are roughly equally spaced. The valley separating peaks four and five includes the tetradecapeptide repeat, which may account for its greater length.

We have also predicted the secondary structure of the E1 polypeptide using two distinct probabilistic methods (Fig. 5 and Discussion). The above peaks of hydrophobicity generally are predicted to be ordered as α-helix or as β-sheet, whereas the valleys are predicted to contain β-turns.

**DISCUSSION**

The Relationship Between E1 and Other Chorion Sequences. One of our major goals in sequencing E1 was to search for possible homologies with other chorion sequences. The lack of homology in the coding and 3' untranslated regions suggests that E1 is not a member of the previously described chorion superfamily, since representative sequences from the two main groups (called A-like and B-like) are more similar to each other than to E1 (Fig. 3). Presumably, the very distinct sequence of E1 is related to its distinct function, which is unlike the functions of the main chorion proteins which assemble into fibrous lamellae.

A feature shared between E1 and other chorion genes is the presence of an intron interrupting the signal peptide-coding region (8, 13). This is not a good indicator of homology, however, as many eukaryotic genes that encode secretory proteins contain introns within the signal peptide-coding region (26).

Another interesting feature shared by chorion genes is the presence of one or more copies of similar hexanucleotide sequences (consensus = ATTCAG) located within the short 5' untranslated region (Fig. 4). In some cases, variants of the consensus sequence are found within a few nucleotides of the cap site, suggesting a possible role in transcriptional initiation. Several sea urchin histone genes also contain 1-nucleotide variants of ATTCAG at their cap sites (27). To see how significant the recurrence of the ATTCAG sequence is in chorion genes, we searched for its presence throughout these genes (see legend to Fig. 4), including ~3100 nucleotides of coding DNA and 690 nucleotides of 3' untranslated DNA. A perfect match was found only once (in coding DNA).

**FIG. 4.** Variants of a hexanucleotide consensus sequence (ATTCAG) found in the 5' untranslated region of chorion genes. Comparisons between the 5' untranslated region of E1 and six other chorion genes, representing four multigene families (HC-A, HC-B, A, and B), were made. The six genes compared with E1 are HC-A (clone 2132, refs. 8 and 14), HC-B (clone 2574, refs. 8 and 14), late A (clone 18c, ref. 13), late B (clone 40la, ref. 13), middle A (clone 292a, ref. 13), and middle B (clone 10a, ref. 13). Below each nucleotide of the consensus sequence is shown the fraction of all sequences compared which are identical. Locations of the sequences are given in nucleotides from the initiating methionine codon. For comparison, the published sites for transcriptional initiation (see above references) are -32 to -36 (HC-A), -26 to -30 (HC-B), -35 (late A), -30 (late B), -35 (middle A), and -32 (middle B). The site of initiation for the E1 gene has not been determined but is predicted to be between -21 and -24 (see legend to Fig. 2).

**FIG. 5.** Hydropathy plot and secondary-structure predictions of the E1 protein. The average hydrophobicity/hydrophilicity of overlapping nonpeptides (called the hydrophobic index, ref. 22) is displayed in the top panel. Values for each nonapeptide are positioned over the fifth residue. The bars with cross-hatching identify contiguous segments with hydrophobic indices greater than zero. Secondary structures (α, α-helix; β, β-sheet; T, β-turn) are derived from prediction schemes of Garnier et al. (thin line and dots, ref. 23) and Chou and Fasman (thick line, refs. 24, 25). In the scheme of Garnier et al., decision constants for all three forms were set at zero. In the Chou and Fasman scheme, β-turns were predicted when the conformational parameter $P_3 > 1.00$, $P_5$, and $P_6$ and when the products of positional frequencies for adjacent tetrapeptides ($p_4$) > 1.30. Vertical arrows identify the positions of cysteine residues. Horizontal arrows below the amino acid sequence at the bottom identify the position of a duplicated tetradecapeptide sequence. The one-letter amino acid abbreviations are used.
This is to be compared with its occurrence four times out of 214 nucleotides of 5' untranslated DNA. When a mismatch of one nucleotide is allowed, the sequence is found 8 times in coding DNA, 9 times in 3' untranslated DNA, and 13 times in 5' untranslated DNA, for an average occurrence of once every 390, 76, and 19 nucleotides, respectively. The ATCAG sequence is not found in the 5' untranslated region of two other well-characterized silkworm genes that encode structural proteins (sercin and fibroin), although 1-nucleotide variants occur four times out of 449 nucleotides in fibroin (28, 29). Given the statistical nature of identifying the consensus sequence, it is premature to speculate whether it might have some function restricted to the 5' untranslated region.

Higher-Order Structure of the E1 Protein. The hydrophobic plot shown in Fig. 5 reveals an interesting periodicity that may be related to higher-order structure of E1. For lamellar sequences, variations in hydrophy indices of overlapping nonapeptides are neither as dramatic nor as periodic (unpublished observations).

Secondary-structure predictions suggest that the hydrophobic stretches in E1 are arranged as $\alpha$-helices and strands of $\beta$-sheet. Peaks 1, 3, and 4 (beginning from the aminoterminal end) are predicted to contain the strands of $\beta$-sheet, whereas peaks 2 and 5 are jointly predicted to contain the $\alpha$-helix. In this scheme, a stretch of $\alpha$-helix within peak 3. $\beta$-Turns are jointly predicted within three of the four relatively hydrophilic valleys and, with lesser confidence, in the remaining valley as well. Beyond these general features, $\alpha$-helix is jointly predicted in the fourth valley within the repeated tetradecapeptide sequence, and a $\beta$-turn is jointly predicted in the middle of peak four.

The predicted secondary structures for lamellar sequences share certain features with E1 but also differ substantially (30, 31). For example, strands of $\beta$-sheet are quite abundant in lamellar sequences, particularly in the conservative interior or central domain, whereas $\alpha$-helix content is quite low. These predictions have been confirmed experimentally (32, 33). By contrast, in E1 $\alpha$-helix is predicted to be slightly more abundant overall than $\beta$-sheet, although both are relatively abundant. In the central domain of lamellar sequences, it has been suggested that $\beta$-turns separate short strands of $\beta$-sheet in the absence of $\alpha$-helix (34). In E1, $\beta$-turns appear to interrupt stretches of both $\beta$-sheet and $\alpha$-helix.

In scanning electron micrographs, filler has a spongy appearance, whereas lamellar chorion consists of discrete, uniformly dense layers (2, 12). In thin-section transmission electron micrographs, filler appears as a loose network of interconnected fibrils, whereas in early- to middle-stage lamellar chorions, the fibrils are more ordered, are more closely packed, and do not form visible connections with each other (3, 4).

The above data permit the construction of a sketchy model for how E1 might assemble to form filler. The five hydrophobic domains are rich in secondary structure and are good candidates for assembling into the core of the filler ultrastructure visible by transmission electron microscopy. Assembly of these regions could either be intra- or intermolecular. Separating these hydrophobic regions are hydrophilic spacers with hinges (the $\beta$-turns) that might permit a change in the orientation of the hydrophobic regions. Particular orientations of the hydrophobic regions could be stabilized by disulfide bonding of cysteine residues localized near selected hinges. Spacer regions could play a role in intermolecular bonding to other spacers, and/or might hydrogen-bond with water molecules. Hydrophilic spacers might hinder formation of a compact ultrastructure, consistent with the generally loose nature of filler; e.g., filler collapses upon dehydration although mature lamellar chorion is unaffected. By contrast, in lamellar chorion, the relative paucity of charged residues may permit formation of a compact ultrastructure. A more detailed model should be possible once the sequence of the other filler component, the E2 protein, is available.

We thank Drs. James Pestell (Harvard University) and Israel Hanukoglu (Technion, Israel) for providing computer software and thank Dr. Priscilla Wilkins Stevens (Northwestern University) for assistance with secondary-structure predictions. We thank our colleague Mr. Antonis Hatzopoulos for clone AP8 and for helpful discussions. This work was supported by a grant from the National Institutes of Health.